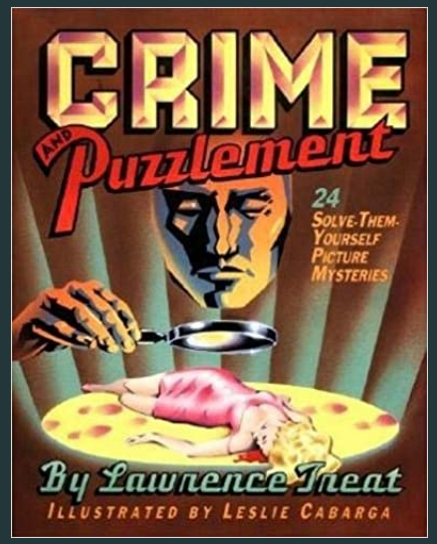
Day 2: Policy evaluation and impact assessment

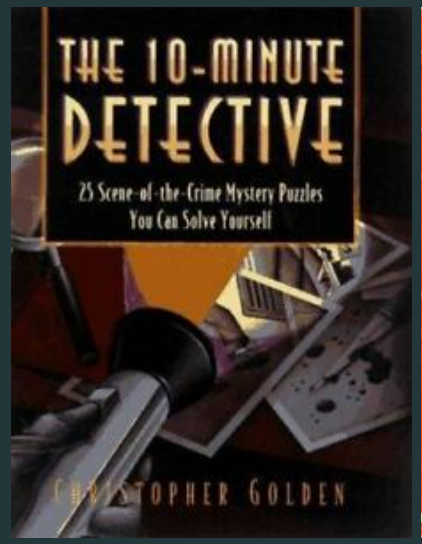
Crimes against causality

Simon Munzert
Hertie School

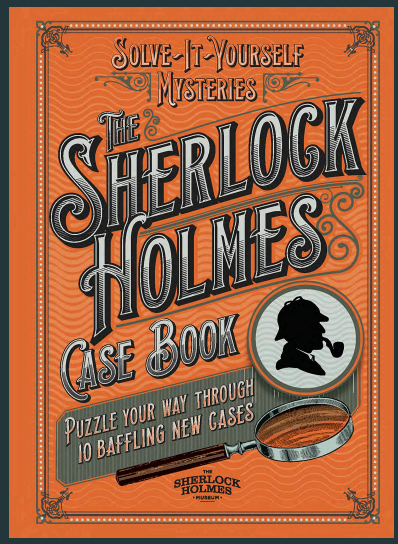
Crimes against causality

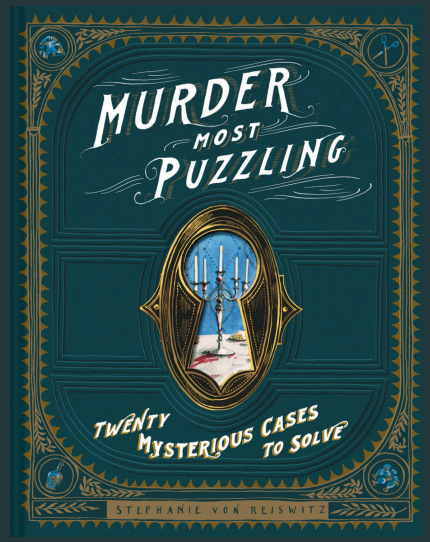

CRIME
AND
Puzzlement
24
SOLVE-THEM-
YOURSELF
PICTURE
MYSTERIES
By Lawrence Treat
ILLUSTRATED BY LESLIE CABARGA

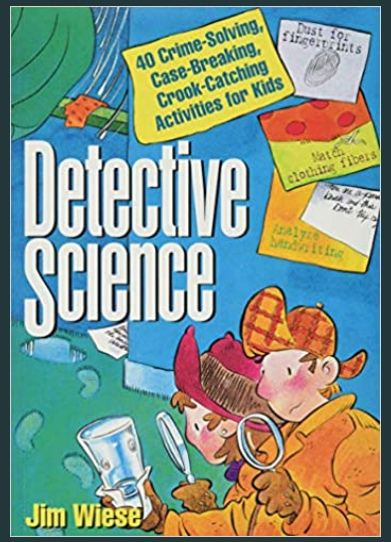

24 Solve-Them-Yourself Picture Mysteries
YOU'RE THE DETECTIVE!
By LAWRENCE TREAT
ILLUSTRATED BY KATHLEEN BOROWIN

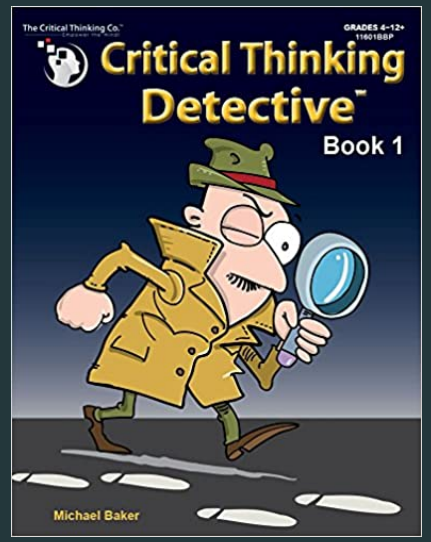

THE 10-MINUTE
DETECTIVE
25 Scene-of-the-Crime Mystery Puzzles
You Can Solve Yourself
CHRISTOPHER GOLDEN

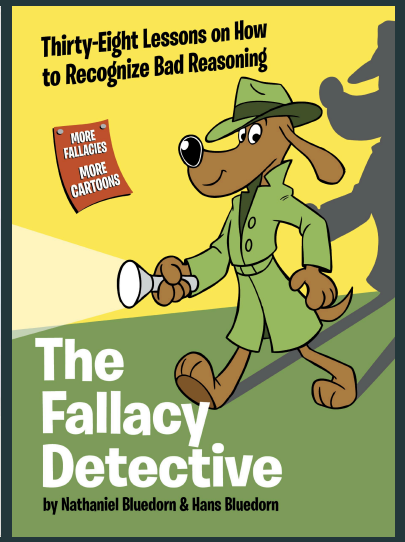

"CRIME CARGO" ··· Crime in Yellow and Scarlet ···
TRUE
DECEMBER
DETECTIVE
MYSTERIES

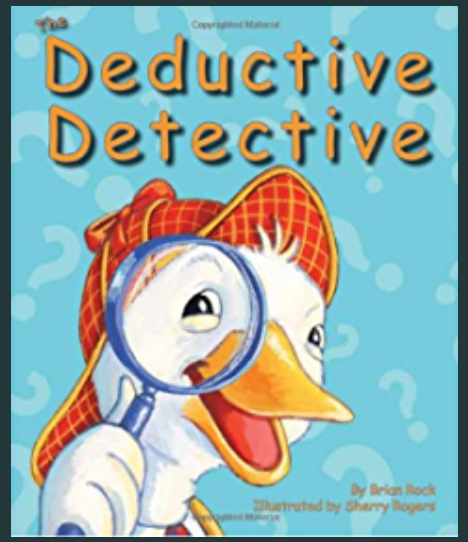

SOLVE-IT-YOURSELF
MYSTERIES
THE
SHERLOCK
HOLMES
CASE BOOK
PUZZLE YOUR WAY THROUGH
10 BAFFLING NEW CASES


MURDER
MOST
PUZZLING
TWENTY
MYSTERIOUS CASES
TO SOLVE
STEPHANIE VON REISWITZ


40 Crime-Solving,
Case-Breaking,
Crook-Catching
Activities for Kids
Detective
Science
Jim Wiese


The Critical Thinking Co.
GRADES 4-12+
Critical Thinking
Detective
Book 1
Michael Baker


Thirty-Eight Lessons on How
to Recognize Bad Reasoning
MORE
FALLACIES
MORE
CARTOONS
The
Fallacy
Detective
by Nathaniel Bluedorn & Hans Bluedorn


Deductive
Detective
By Brian Rock
Illustrated by Sherry Rogers

BOUDOIR

Amy LaTour's body was found in her bedroom last night, as shown, with her pet canary strangled in its cage. Henry Willy and Joe Wonty, her boyfriends; Louis Spanker, a burglar known to have been in the vicinity; and Celeste, her maid, were questioned by the police.

Wilbur Unisex, who happened to be in the area pursuing the *Heliconius charitonius*, put down his butterfly net and solved the case. Can you?

Questions

1. How was Amy apparently killed?
   Shot____ Stabbed____
   Strangled____ Beaten____
2. Is there evidence of a violent struggle? Yes____ No____
3. Was her murderer strong?
   Yes____ No____
4. Was Amy fond of jewelry?
   Yes____ No____
5. Was she robbed? Yes____ No____
6. Do you think she had been on friendly terms with her killer?
   Yes____ No____
7. Was the canary strangled before Amy's death? Yes____ No____
8. Was this a crime of passion?
   Yes____ No____
9. Did Willy have a motive?
   Yes____ No____
10. Who killed Amy? Henry Willy____
    Joe Wonty____ Louis Spanker____
    Celeste____

Solution on page 51

What those books are about

Critical Thinking Detective

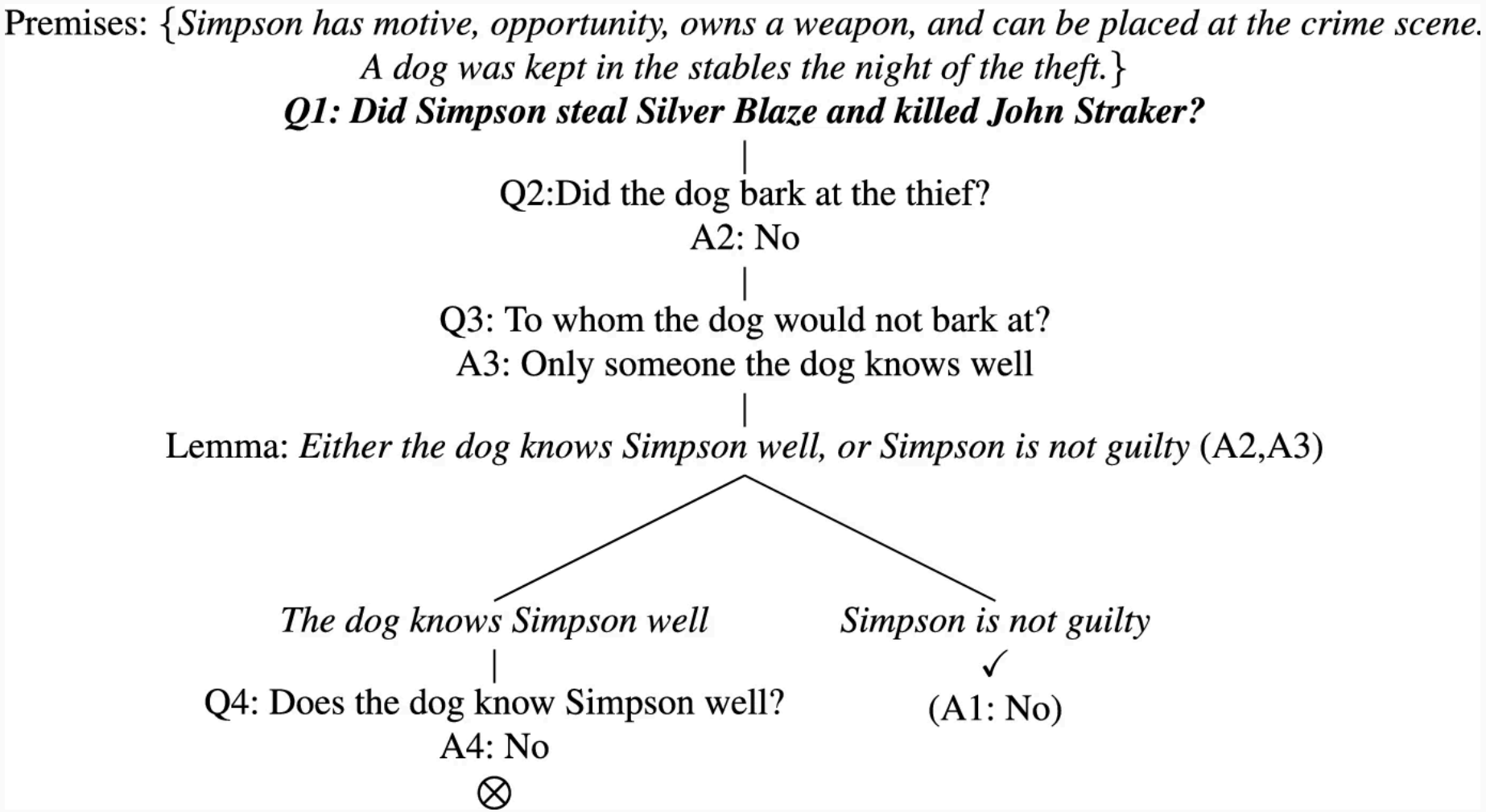
"This colorful 32-page book offers a collection of fun, easy-to-use detective cases for Grades 4–12+. Some cases may be more challenging for younger students, but teachers and parents can always provide hints when needed. The cases develop critical thinking skills by requiring students to read carefully and analyze and synthesize information to guide their decision-making. The cases also develop observation skills, reading comprehension, deductive and inductive thinking skills. The ability to identify and evaluate evidence is the very heart of critical thinking."

The Sherlock Holmes Case Book

"Join the world's greatest fictional detective and use your own powers of deduction to solve these puzzling mysteries. The Sherlock Holmes Case Book: Solve-It-Yourself Mysteries is a remarkable collection of crimes from Dr John Watson's case notes that features all of the twists and turns that have come to be expected from a Holmes case – but now it is up to you to solve them.

There are 10 cases to be cracked, each of which requires the reader to use logic and powers of perception to answer a question at three points – the beginning, the middle and the end."

The Holmesian "science of deduction and analysis"



The Holmesian "science of deduction and analysis"

> "[F]rom a drop of water [...] a logician could infer the possibility of an Atlantic or a Niagara without having seen or heard of one or the other [...]. By a man's finger-nails, by his coat-sleeve, by his boot, by his trouser-knees, by the callosities of his forefinger and thumb, by his expression, by his shirt-cuffs—by each of these things a man's calling is plainly revealed. That all united should fail to enlighten the competent inquirer in any case is almost inconceivable."

Arthur Conan Doyle, *A Study in Scarlet*

> "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?"

Arthur Conan Doyle, *The Sign of the Four*

↪ **Great literature, but a nonsensical strategy for causal reasoning.**

Causal thinking as detective work

The premise

- Observational **causal inference** is like detective work[1]
- We have to **piece together evidence** to solve a puzzle of whether the suspect - the treatment - in fact causally changed the outcome
- To that end, we need to **eliminate alternative explanations**

The equipment

- **Our toolbox:** Critical thinking and logic, observation (data collection), perception (measurement), external knowledge, and statistical methods
- Without the qualitative skills, the quantitative methods are pointless

[1] "Observational" does some heavy lifting here; for the experimental causal inference analogy, we'd probably the ones committing the murder.

Warming-up

A few easy cases

Shoes hurt (?)

> Survey data revealed that people who sleep with their shoes on are much more likely to wake up with a headache.

Sweetened beverages are fattening (?)

> Individuals regularly consuming sugar sweetened beverages are shown to have a 30% higher BMI.

Lock-downs caused deaths due to COVID-19 (?)

> European countries that had stricter & longer lock-down periods had a higher COVID-19 death rate.

UN missions fail to protect civilians (?)

> UN peace keeping missions in civil war scenarios are strongly associated with higher death rates among civilians.

Causal crime scene #1

The scene: Do storks deliver babies?

Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

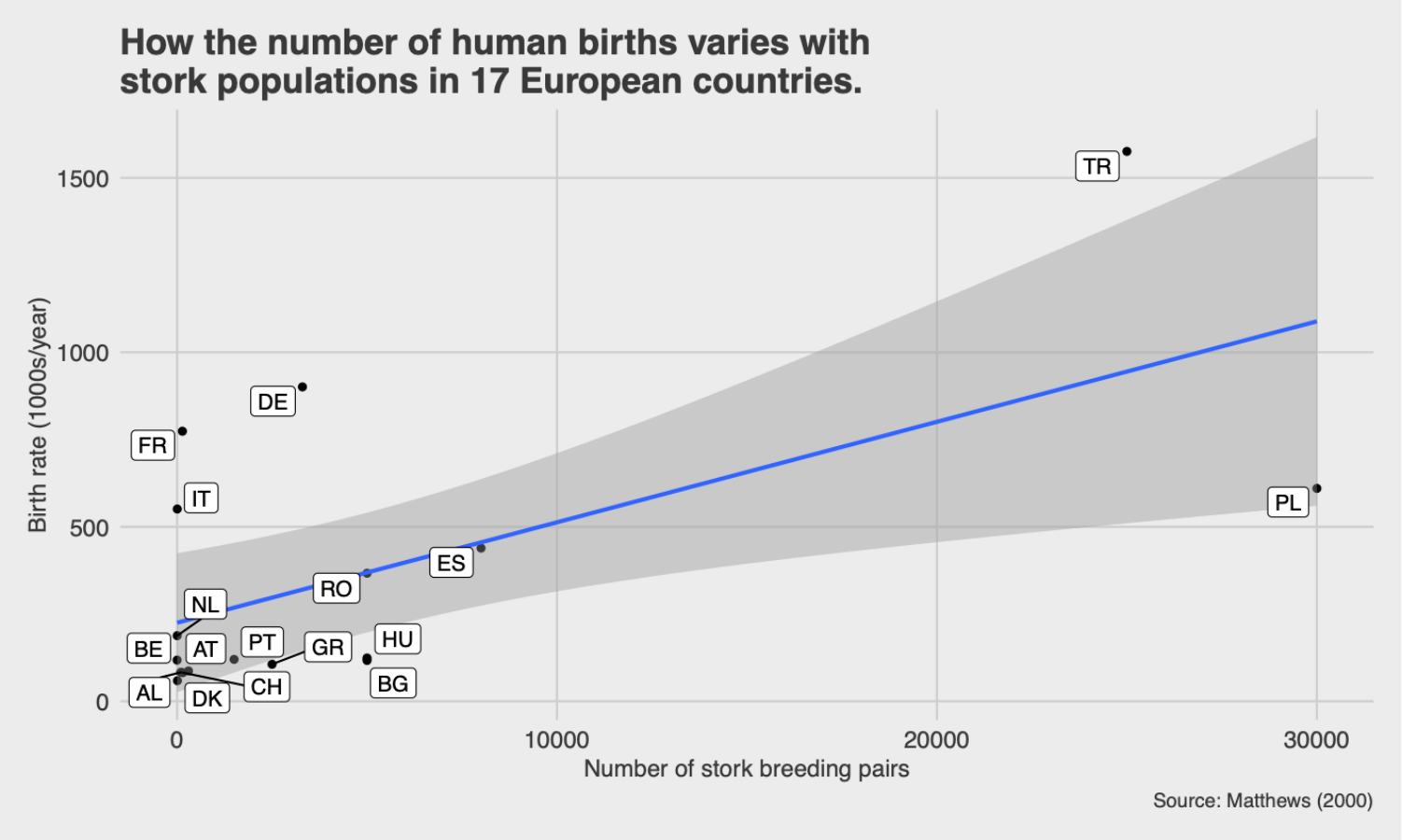
◆ INTRODUCTION ◆

Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high corre-

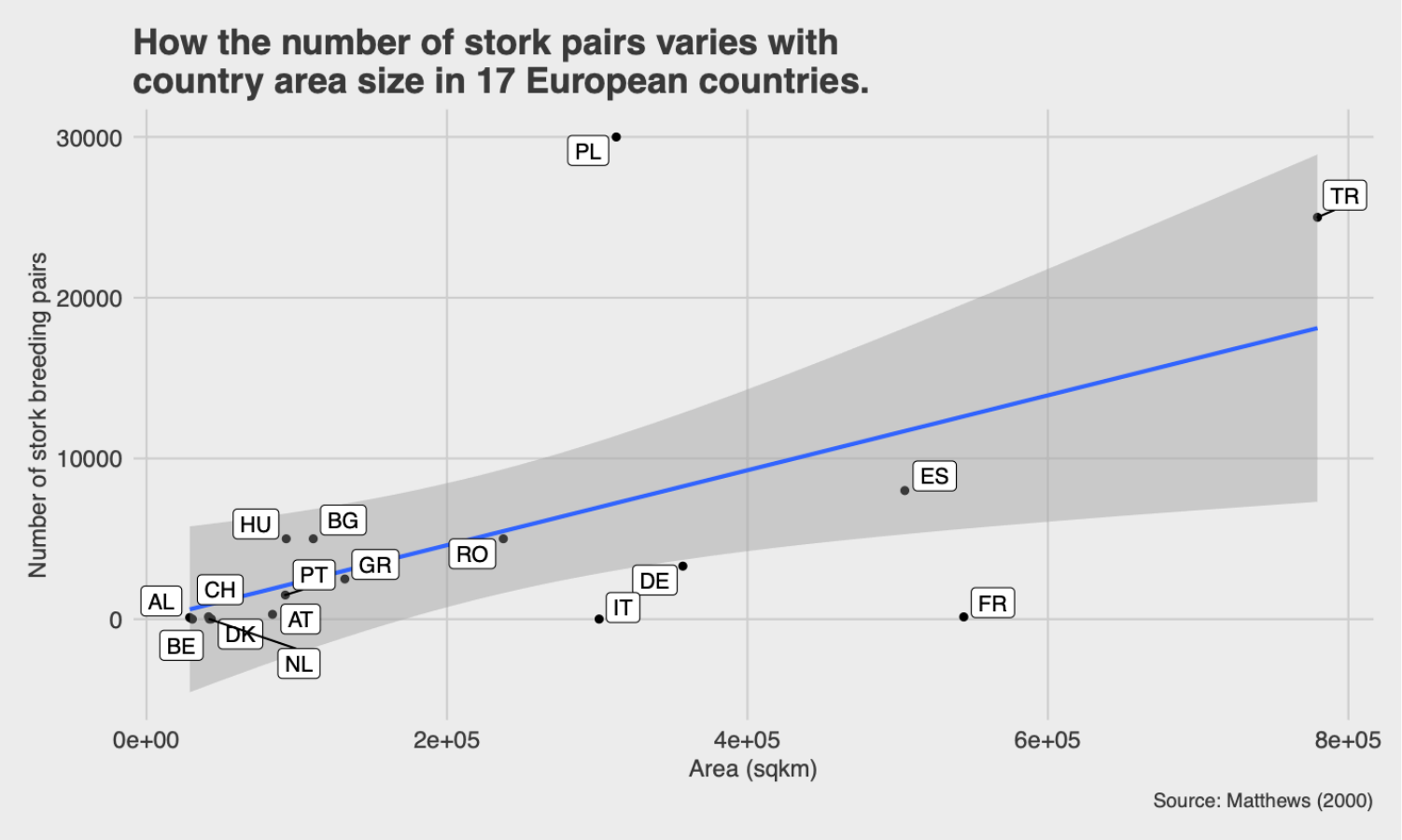
association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings

`Source` Matthews, Roger, 2000, Teaching Statistics

Additional evidence

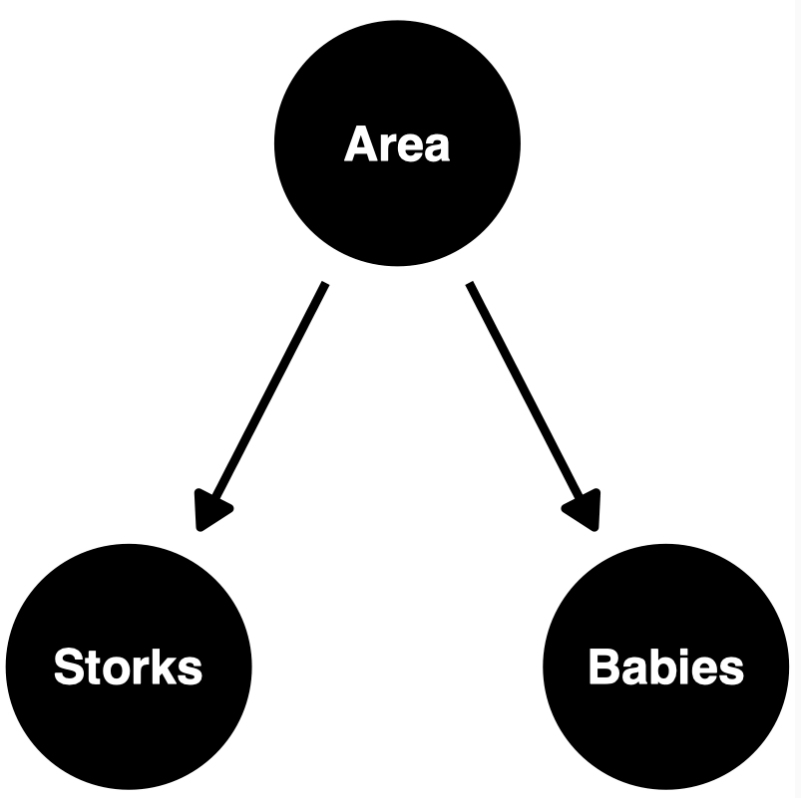


Additional evidence



A solution to the case

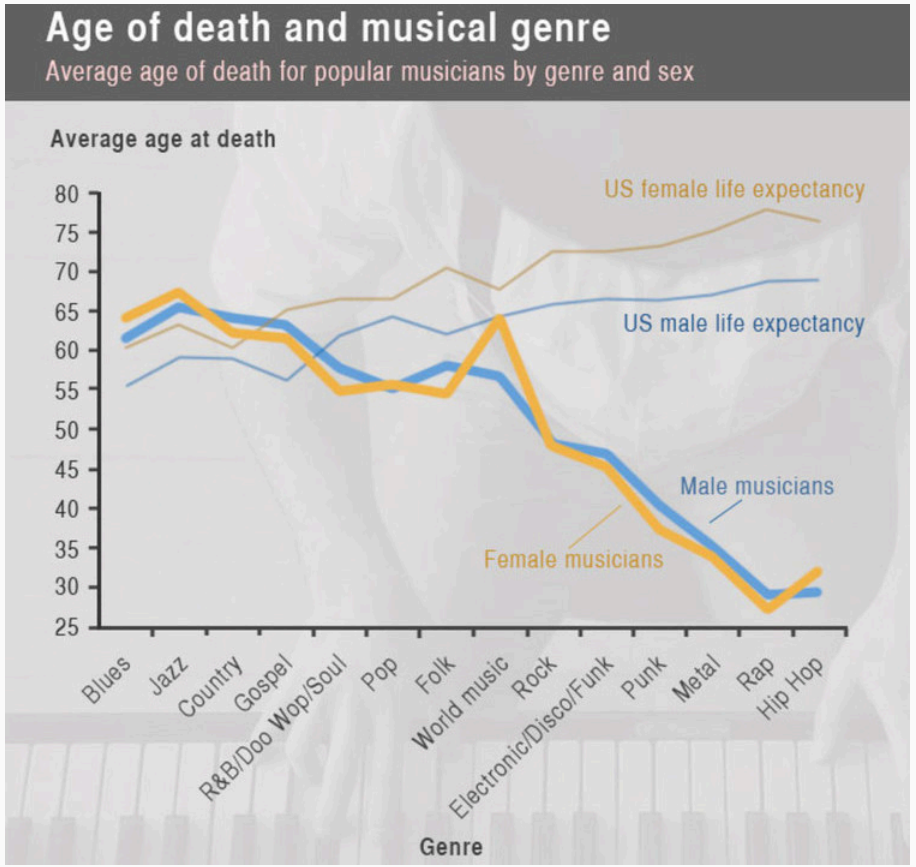
	Evidence and conclusions
The crime scene	Storks correlate with birth rates
The suspect(s)	Area, breeding grounds, urban/rural divide, chance
The murder weapon	Confounding



Causal crime scene #2

The scene: Are certain music genres deadlier than others?

Hertie School



`Source` The Conversation, `Background (w/ author response)` Calling Bullshit

> [Study author Dianna Kenny] found that musicians from older genres – including blues, jazz, country and gospel – have similar lifespans to American people their own age. The life expectancy for R&B musicians is slightly lower, while the life expectancy for newer genres like rock, techno, punk, metal, rap and hip hop is significantly shorter."
>
> Ana Swanson, Washington Post

> "It's a cautionary tale to some degree," Kenny told the Washington Post. "People who go into rap music or hip hop or punk, they're in a much more occupational hazard profession compared to war. We don't lose half our army in a battle."
>
> Dianna Kenny, quoted by Washington Post

Additional evidence

Cause of death by genre

Various causes of death for musicians of different genres

	Accidental	Suicide	Homicide	Heart-related	Cancer
% deaths per cause	19.5%	6.8%	6.0%	17.4%	23.4%
Blues	9.2%	2.0%	3.5%	28.0%	24.2%
Jazz	10.6%	2.7%	1.9%	20.7%	30.6%
Country	15.8%	4.7%	1.6%	23.5%	25.1%
Gospel	13.3%	0.9%	3.6%	18.5%	23.0%
R&B	11.5%	1.6%	5.0%	23.2%	26.8%
Pop	19.0%	6.4%	2.9%	16.4%	26.7%
Folk	15.9%	5.5%	4.4%	15.3%	32.3%
World music	12.7%	3.4%	9.6%	17.8%	19.9%
Rock	24.4%	7.2%	3.6%	15.4%	24.7%
Electronic	16.7%	5.0%	10.0%	15.0%	25.0%
Punk	30.0%	11.0%	8.2%	12.6%	18.3%
Metal	36.2%	19.3%	5.9%	11.0%	14.1%
Rap	15.9%	6.2%	51.0%	6.9%	7.6%
Hip Hop	18.3%	7.4%	51.5%	6.1%	6.1%

Red: significantly above the overall average rate for cause of death
Blue: above the overall average rate for cause of death
Green: significantly below the overall average rate for cause of death

Note: not all causes shown

theconversation.com

Source: Author

Some issues

1. **Sanity check:** Are some genres really that deadly? E.g., do rap musicians really die at an average age of ~30?
2. **Right censoring:** Some genres are younger than others and the data are conditional on musicians having died already. Most rap and hip-hop stars are still alive today; we don't know how long they'll live!
3. **Conditional probabilities:** The probabilities of each cause of death are conditional on death having already occurred at the time of the study.
4. (Minor issue) A line graph for categorical data? Not a good idea.

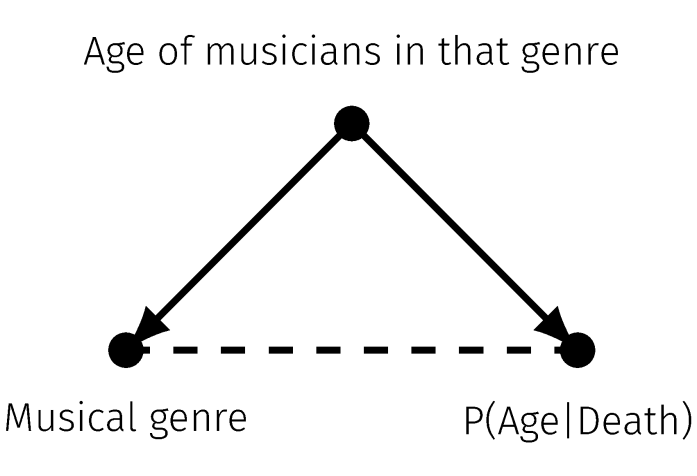
A solution to the case

	Evidence and conclusions
The crime scene	Musical genre correlates with age at death
The suspect(s)	Life style (e.g. drug consumption), age of genre, incomplete data collection
The murder weapon	Confounding via data censoring

> "In other words, it's not that rap stars will likely die young; it's that the rap stars who have died certainly died young because rap hasn't been around long enough for it to be otherwise."
>
> Carl Bergstrom and Jevin West, Calling Bullshit[1]

Inspector Lestrade:

Sherlock:



[1]Note that (1) the author, Dianna Kenny, has provided a nice and plausible response and that (2) a more rigorous study seems to provide evidence consistent with the original patterns.

Causal crime scene #3

The scene: Is air pollution affecting life expectancy?

Hertie School

Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

Yuyu Chen, Avraham Ebenstein, Michael Greenstone, and Hongbin Li Authors Info & Affiliations

Edited by William C. Clark, Harvard University, Cambridge, MA, and approved May 28, 2013 (received for review January 2, 2013)

July 8, 2013 | 110 (32) 12936-12941 | https://doi.org/10.1073/pnas.1300018110

Abstract

This paper's findings suggest that an arbitrary Chinese policy that greatly increases total suspended particulates (TSPs) air pollution is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The quasi-experimental empirical approach is based on China's Huai River policy, which provided free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. Using a regression discontinuity design based on distance from the Huai River, we find that ambient concentrations of TSPs are about 184 μg/m^3 [95% confidence interval (CI): 61, 307] or 55% higher in the north. Further, the results indicate that life expectancies are about 5.5 y (95% CI: 0.8, 10.2) lower in the north owing to an increased incidence of cardiorespiratory mortality. More generally, the analysis suggests that long-term exposure to an additional 100 μg/m^3 of TSPs is associated with a reduction in life expectancy at birth of about 3.0 y (95% CI: 0.4, 5.6).

Table 3. Using the Huai River policy to estimate the impact of TSPs (100 μg/m^3) on health outcomes

Dependent variable	(1)	(2)	(3)
Panel 1: Impact of "North" on the listed variable, ordinary least squares			
TSPs, 100 μg/m^3	2.48*** (0.65)	1.84*** (0.63)	2.17*** (0.66)
ln(All cause mortality rate)	0.22* (0.13)	0.26* (0.13)	0.30* (0.15)
ln(Cardiorespiratory mortality rate)	0.37** (0.16)	0.38** (0.16)	0.50*** (0.19)
ln(Noncardiorespiratory mortality rate)	0.00 (0.13)	0.08 (0.13)	0.00 (0.13)
Life expectancy, y	−5.04** (2.47)	−5.52** (2.39)	−5.30* (2.85)
Panel 2: Impact of TSPs on the listed variable, two-stage least squares			
ln(All cause mortality rate)	0.09* (0.05)	0.14** (0.07)	0.14* (0.08)
ln(Cardiorespiratory mortality rate)	0.15** (0.06)	0.21** (0.09)	0.23** (0.10)
ln(Noncardiorespiratory mortality rate)	0.00 (0.05)	0.04 (0.07)	0.00 (0.06)
Life expectancy, y	−2.04** (0.92)	−3.00** (1.33)	−2.44 (1.50)
Climate controls	No	Yes	Yes
Census and DSP controls	No	Yes	Yes
Polynomial in latitude	Cubic	Cubic	Linear
Only DSP locations within 5° latitude	No	No	Yes

The sample in columns (1) and (2) includes all DSP locations (n = 125) and in column (3) is restricted to DSP locations within 5° latitude of the Huai River boundary (n = 69). Each cell in the table represents the coefficient from a separate regression, and heteroskedastic-consistent SEs are reported in parentheses. Models in column (1) include a cubic in latitude. Models in column (2) additionally include demographic and climate controls reported in Table 1. Models in column (3) are estimated with a linear control for latitude. Regressions are weighted by the population at the DSP location. *Significant at 10%, **significant at 5%, ***significant at 1%. Sources: China Disease Surveillance Points (1991–2000), *China Environment Yearbook* (1981–2000), and World Meteorological Association (1980–2000).

Additional evidence

Hertie School

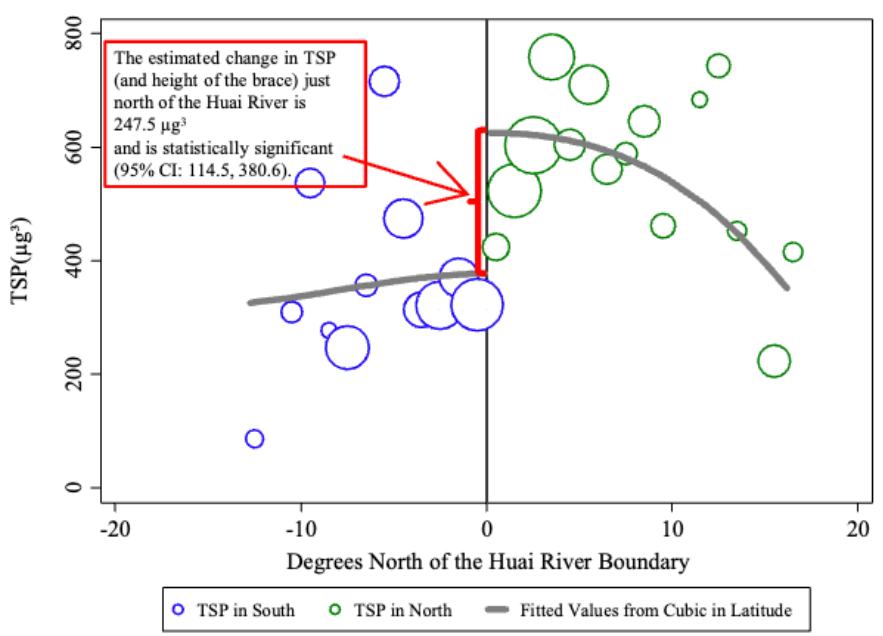


Fig. 2. Each observation (circle) is generated by averaging TSPs across the Disease Surveillance Point locations within a 1° latitude range, weighted by the population at each location. The size of the circle is in proportion to the total population at DSP locations within the 1° latitude range. The plotted line reports the fitted values from a regression of TSPs on a cubic polynomial in latitude using the sample of DSP locations, weighted by the population at each location.

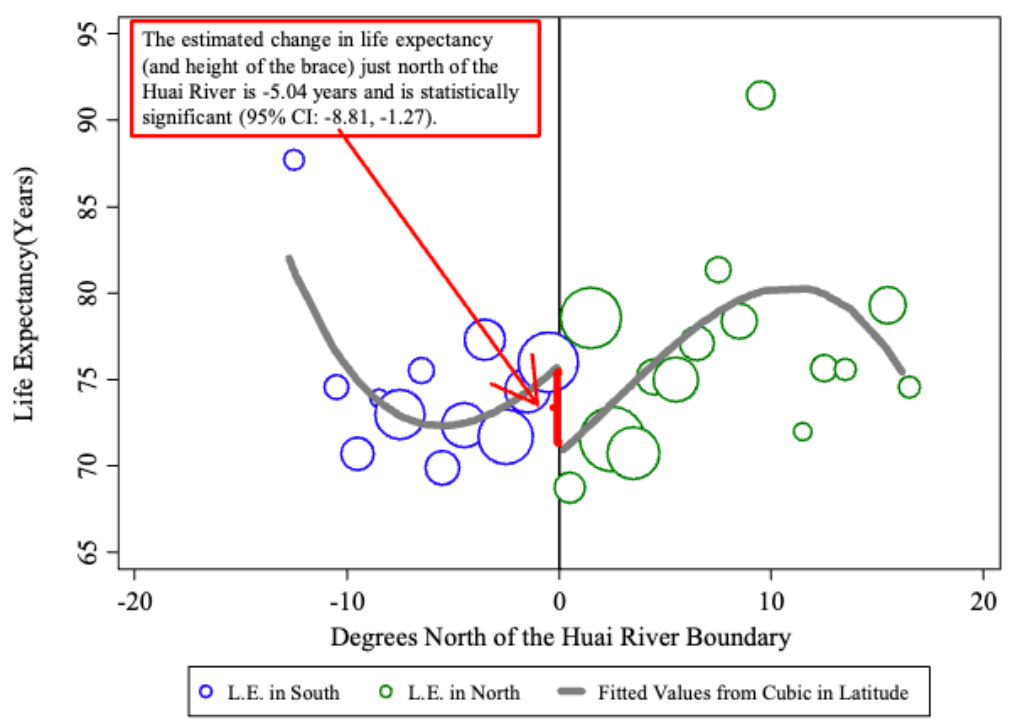


Fig. 3. The plotted line reports the fitted values from a regression of life expectancy on a cubic in latitude using the sample of DSP locations, weighted by the population at each location.

Additional evidence (from a later study)

New evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River Policy

Avraham Ebenstein, Maoyong Fan, Michael Greenstone, +1, and Maigeng Zhou Authors Info & Affiliations

Edited by William C. Clark, Harvard University, Cambridge, MA, and approved July 3, 2017 (received for review October 27, 2016)

September 11, 2017 | 114 (39) 10384-10389 | https://doi.org/10.1073/pnas.1616784114

Abstract

This paper finds that a 10-μg/m^3 increase in airborne particulate matter [particulate matter smaller than 10 μm (PM_{10})] reduces life expectancy by 0.64 years (95% confidence interval = 0.21–1.07). This estimate is derived from quasiexperimental variation in PM_{10} generated by China's Huai River Policy, which provides free or heavily subsidized coal for indoor heating during the winter to cities north of the Huai River but not to those to the south. The findings are derived from a regression discontinuity design based on distance from the Huai River, and they are robust to using parametric and nonparametric estimation methods, different kernel types and bandwidth sizes, and adjustment for a rich set of demographic and behavioral covariates. Furthermore, the shorter lifespans are almost entirely caused by elevated rates of cardiorespiratory mortality, suggesting that PM_{10} is the causal factor. The estimates imply that bringing all of China into compliance with its Class I standards for PM_{10} would save 3.7 billion life-years.

Table 2. RD estimates of the impact of the Huai River Policy

Outcome	[1]	[2]	[3]
Pollution and life expectancy			
PM_{10}	27.4*** (9.5)	31.8*** (9.1)	41.7*** (12.9)
Life expectancy at birth, y	−2.4** (1.0)	−2.2* (1.1)	−3.1*** (0.9)
Cause-specific mortality (per 100,000, log)			
Cardiorespiratory	0.30** (0.14)	0.22* (0.13)	0.37*** (0.11)
Noncardiorespiratory	0.06 (0.10)	0.08 (0.09)	0.13 (0.08)
RD type	Polynomial	Polynomial	LLR
Polynomial function	Third	Linear	
Sample	All	5°	

Column [1] reports OLS estimates of the coefficient on a north of the Huai River dummy after controlling for a polynomial in distance from the Huai River interacted with a north dummy using the full sample (n = 154) and the control variables from *SI Appendix*, Table S1. Column [2] reports this estimate for the restricted sample (n = 79) of DSP locations within 5° of the Huai River. Column [3] presents estimates from local linear regression (LLR), with triangular kernel and bandwidth selected by the method proposed by Imbens and Kalyanaraman (14).
*Significant at 10%.
**Significant at 5%.
***Significant at 1%.

Additional evidence (from a later study)

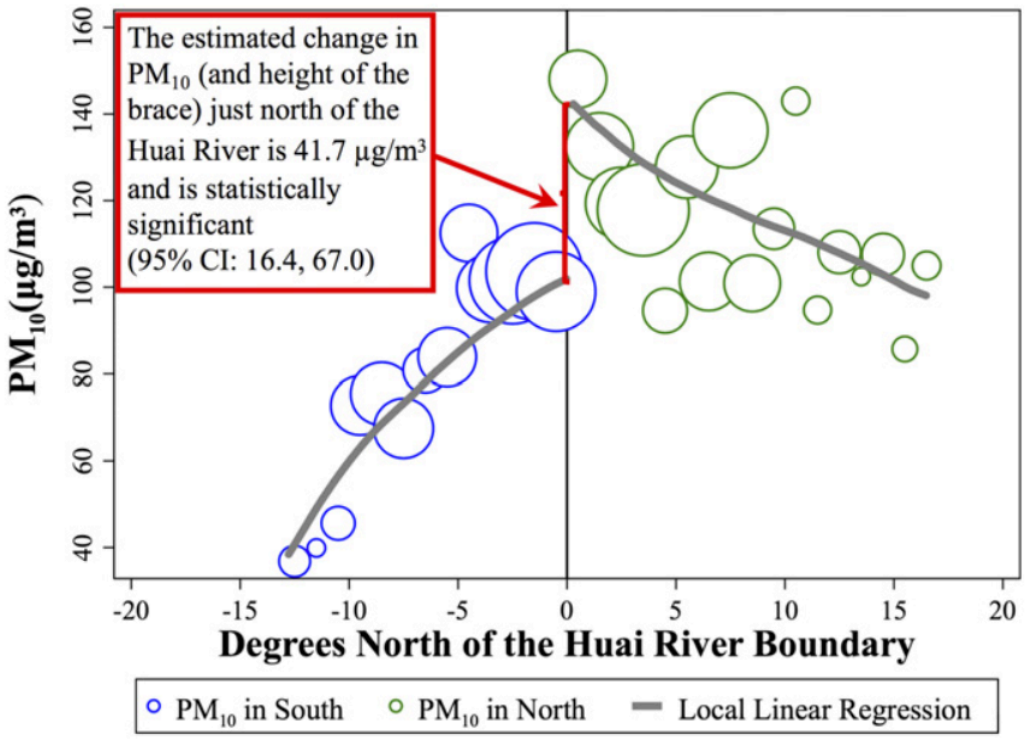


Fig. 2. Fitted values from a local linear regression of PM_{10} exposure on distance from the Huai River estimated separately on each side of the river.

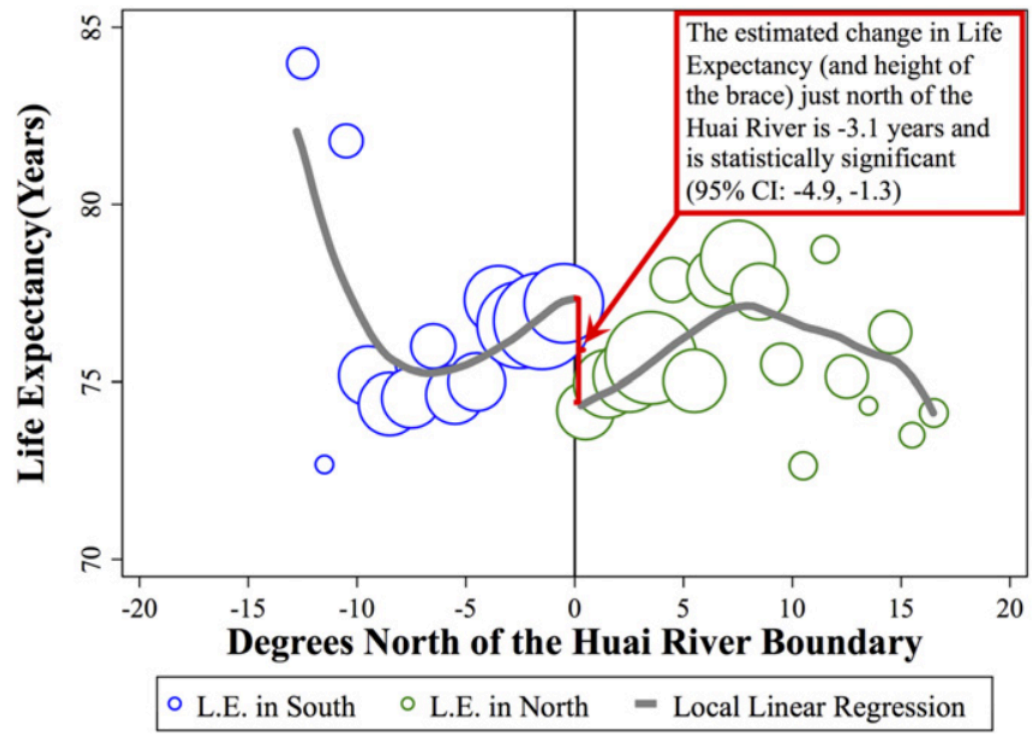


Fig. 3. Fitted values from a local linear regression of life expectancy (L.E.) on distance from the Huai River estimated in the same manner as in Fig. 2.

A solution to the case

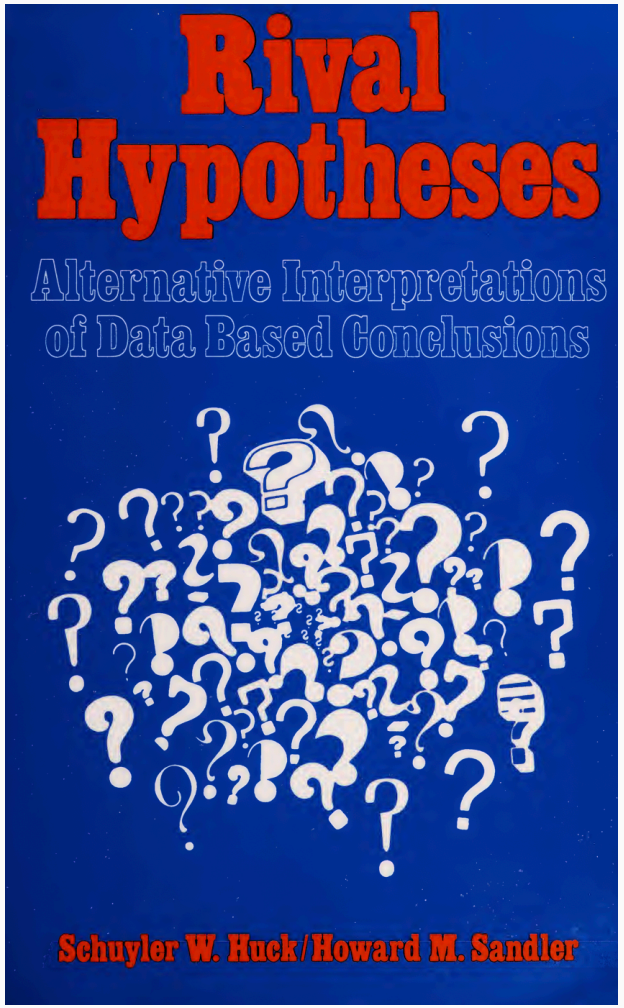
	Evidence and conclusions
The crime scene	Living north versus south of river Huai linked to lower LE
The suspect(s)	Quasi-experimental exposure to PM_{10}
The murder weapon	Likely PM_{10}?

Some thoughts

1. A quasi-experimental design does not protect you from measurement or model specification issues.
2. Using higher-order polynomials in RDDs can be problematic and generate bias (see Gelman and Zelizer, 2015.
3. The follow-up study seemed to address many of those issues.
4. Science (sometimes) is a self-correcting enterprise. Progress through higher quality of evidence is possible.

More causal murders to be discovered

Hertie School



20 Categories of Rival Hypotheses (And a Classification of the 100 Problems)

(1)	*Correlation Causality*	(2)	*Cross-Sectional/ Longitudinal*	(3)	*Experimenter Effect*	(4)	*History*	(5)	*Instability*
	3, 9, 37, 75, 77, 83, 91, 95, 100		22, 55, 89		7, 16, 21, 28, 78, 98		5, 19, 23, 30, 39, 41, 66, 69, 80, 81, 97		8, 31, 63, 65, 94
(6)	*Instrumentation*	(7)	*Matching*	(8)	*Maturation*	(9)	*Mortality*	(10)	*Observer/Rater Effect*
	56, 60, 69, 80, 81		35, 84, 95		47, 48, 81		2, 16, 22, 27, 42, 79, 81, 99		34, 40, 56, 99
(11)	*Order Effect*	(12)	*Regression*	(13)	*Sampling Bias*	(14)	*Selection*	(15)	*Statistical, Other Than Instability and Regression*
	21, 31, 44, 66, 97		11, 35, 59, 81		18, 44, 51, 61, 63, 66, 71, 89, 93		22, 29, 30, 49, 53, 65, 67, 72, 73, 85, 98		20, 53, 64, 67, 72, 78, 89
(16)	*Subject Effect*	(17)	*Testing*	(18)	*Treatment Confound*	(19)	*Valid Data/ Self-Report*	(20)	*Different Interpretation of Data*
	25, 26, 30, 31, 38, 58, 61, 62, 65, 68		5, 13, 23, 80, 81		1, 6, 7, 12, 14, 15, 16, 17, 18, 25, 28, 32, 37, 45, 52, 66, 70, 73, 74, 78, 87, 92, 98		9, 19, 25, 36, 48, 57, 65, 77		8, 18, 88, 96

More causal murders to be discovered

Hertie School

(1) Correlation and Causality

If a group of people or objects is measured with respect to two variables, and if neither of the variables is experimentally manipulated, then the simple finding of a direct or indirect relationship between the two variables—through some sort of correlation coefficient (r_{pb} or 0) or statistical test (t or X^2)—cannot serve as legitimate evidence for the claim that one variable has a causal influence on the other. Quite possibly, both of the variables are causally linked to some unmeasured third variable.

(2) Cross-Sectional/Longitudinal

Researchers are often interested in identifying developmental trends in people or animals—that is, physiological or psychological changes that are simply a product of aging. One popular strategy involves measuring, at one point in time, subjects who differ widely in age, and then comparing the age subgroups on the variable(s) of interest to infer the developmental trend. The other simple strategy involves measuring a single group of subjects repeatedly as they age, with developmental inferences drawn from an observation of how the group's performance varies with time. Unfortunately, neither of these strategies is able to assess validly the developmental process. In the cross-sectional approach, subgroup differences may have been brought about by differential lengths of exposure to improving (or worsening) environmental or nutritional conditions, not by age differences. In the longitudinal approach, changes over time might well be tied to nonmaturational events that are unique to that one group's life span.

(3) Experimenter Effect

If there are two or more conditions (levels) of a manipulated treatment variable, and if the treatments are administered to the subjects by someone familiar with the researcher's hypotheses and hoped-for results, then it is possible that the comparison groups will be treated differently along dimensions other than that associated with the formal independent variable. In nonexperimental descriptive studies wherein comparisons are made between the way two or more status groups (such as males and females) react to the same stimulus, the same sort of bias can exist if the stimulus is being presented to the subjects by a person who has a preference (possibly unconscious) that the groups show up as dissimilar.

(4) History

When a group of subjects is measured before and after exposure to some sort of treatment (or nonexperimental activity or event), a pretest–post-test change

or lack of change in the data collected may be attributable to something other than the treatment and that took place outside the confines of the experiment between the pre- and post-test measurements. Clearly, history can make a treatment look as if it made a difference when in fact the treatment was inert. Or, worse yet, a truly beneficial (or detrimental) treatment may end up looking as if it was detrimental (or beneficial) if its effect is confounded with a negative (or positive) historical event.

(5) Instability

Instability refers to the fact that sample statistics almost never turn out to be the same as the corresponding population parameters, even though the sample is chosen so that its estimates of the population values are unbiased. If two treatments having either the same effect or no effect were applied to groups of randomly formed subjects, we would expect the sample means (or other statistics) to differ. When expected differences are neglected by a failure to set up levels of significance and statistical tests, the researcher ends up drawing conclusions that are almost certain to constitute Type I errors whenever H_o is true and Type III errors about half the time when H_o is false by a small amount.

(6) Instrumentation

Sometimes a measuring instrument's ability to yield accurate data changes in a systematic manner over time, as when the norms of a standardized test gradually (or sometimes quickly) become obsolete. It is as if the springs of a scale had become stretched. If such an instrument is being used to measure a single group of people so as to provide pretest and post-test data, the change in the measuring instrument may make it appear as if the treatment had more or less of an impact than was really the case. The problem of instrumentation might also arise in post-test–only designs if all the members of one group are measured before any members of other groups and the instrument is affected by use (such as a scale). Another likely situation in which this rival hypothesis becomes confounded with treatment effects occurs when observers or raters are required to use a complex recording form, when there are different people doing the recording at pre and post (or for each group in a multigroup study), or when practice effects or boredom leads to a change (over time) in the recorder's ability to use the instrument in the proper manner.

(7) Matching

In comparing two or more treatments against one another (and possibly against a control condition), researchers often use the technique of matching

to decide which subjects will be exposed to the treatment conditions, or which from among a large pool that received the treatment will be measured and/or have their data analyzed. Sometimes the researcher uses this technique because it is impossible to assign subjects randomly to the various comparison conditions; at other times the researcher is probably under the impression that matching works as well as or better than randomization. Regardless of the reasons for its use, the technique of matching does not ensure that the comparison groups are equivalent. While it does rule out the possibility that the treatment variable will be confounded with group differences on the variable(s) used to do the matching, there still remains the possibility that one or more of the nearly infinite number of variables *not* used in the matching process is more related to the obtained group differences than the treatment variable. And as we explain carefully in the solution to problem 35, matching can also bring forth the phenomenon of regression towards the mean.

(8) Maturation

Many characteristics of humans, animals, and plants change over time as a natural consequence of internal events associated with the aging process. If pretest and post-test data are collected on a single group of experimental subjects, a pre-post change may be attributable to the intervening treatment that was administered by the researcher. However, in many such studies it is impossible to rule out the possiblity that maturation has caused a beneficial (or detrimental or inert) treatment to look as if it had an impact different from what was actually the case.

(9) Mortality

If subjects drop out of a one-group pretest–post-test design, or if there are differential rates of (and reasons for) attrition in multigroup designs, conclusions regarding treatment effects may be misleading. As is the case with other rival hypotheses, mortality may cause a truly beneficial treatment to look worthless or detrimental, a truly detrimental treatment to look worthless or beneficial, or an inert treatment to look as if it has had a good or bad impact.

(10) Observer/Rater Effect

In some research studies, people observe the subjects and record their reactions to the treatments that have been administered. In other studies, people rate audiotapes, videotapes, or photographs of the subjects or some written documents (like essay tests) or other tangible items produced by the subjects while under the possible influence of the treatments. In either situation, the conclusions of the study could be misleading if the observers or the raters were

A checklist for policy supporting research

Hertie School

Characteristic to be checked	Quality dimensions	Most likely location in the report
1. Is it a research question?	knowing vs. prescribing	introduction, conclusion
2. Is the research question answerable?	suitability for empirical research	introduction, conclusion, executive summary
3. What kind of knowledge is needed?	data produced by design meets data required by question (descriptive, exploratory, confirmatory)	introduction, methods, conclusion
4. What order of data is required?	order of data secured (real world, experience, research context) is that required to answer question inferences justified	introduction, methods, analysis

Characteristic to be checked	Quality dimensions	Most likely location in the report
5. What level of data is required?	level of research (sub-individual individual, collective) = level of claim inferences justified	introduction
6. What quality of data is required?	appropriate design accounts for field compromises	methods, limitations
7. What methods of analysis are required?	adequately discussed appropriate	methods, analysis, limitations
8. Do the research results support the conclusions?	sound inference accounts for threats	analysis, limitations
9. Do the conclusions provide an answer to the research question?	equivalence	introduction, conclusion

6 Closing remarks: applying the checklist in practice

In this essay we have outlined a (fast and simple) protocol that can be used by policymakers to decide if they should reject a report of empirical social science research.[11] The first criterion tested if the report was structured around a researchable question, the second if the question was answerable, the third tested for the type of knowledge claim required and the fourth proposed a number of questions to test the empirical foundations of the report. Sequentially, the most efficient procedure to quickly assess a research report is by taking the following steps:

1. First check whether the research question is proper, that is aimed at gaining knowledge, not at changing reality (if not, dismiss report on the ground that it cannot involve research).
2. Then check whether it is potentially answerable or not (if not, dismiss report on the ground that it cannot be researched).
3. Then check whether or not there is a mismatch between the conclusion and the research question (if there is a mismatch, dismiss report on the ground that the research commission has not been fulfilled).
4. When there is no apparent mismatch between the conclusion and the CRQ, check if there is a mismatch between the conclusions and the empirical research findings (if so, dismiss report on the ground that the conclusions are not substantiated empirically).
5. When there is no apparent mismatch between the results and the conclusions, check for a mismatch between study design and the kind of knowledge required to answer the RQ (if there is a mismatch, dismiss report on the ground that research is badly designed).
6. If study design and kind of knowledge are compatible, check that the order and level of data used in the study match those required to formulate the conclusion (if there is a mismatch, dismiss report on the ground that it cannot substantiate claims empirically).
7. Check whether the data collecting process and its analysis is fully documented (if incomplete or unclear, dismiss report on the ground that its data cannot be trusted and/or its methods are not transparent).
8. If the report does not immediately fail on one of the counts 1–7 listed above, more careful study of the document is in order (still keeping in mind the checklist of Table 1).

Source: Kamen and Tamás, 2014, "Should I take this seriously? A simple checklist for calling bullshit on policy supporting research", Quality & Quantity